Progress on the PORTIA Project

JOAN FEIGENBAUM
http://www.cs.yale.edu/homes/jf
March 21, 2005; Rutgers

PORTIA: Privacy, Obligations, and Rights in Technologies of Information Assessment

Large-ITR, five-year, multi-institutional, multi-disciplinary, multi-modal research project on end-to-end handling of sensitive information in a wired world

http://crypto.stanford.edu/portia/

Motivation

- Sensitive Information: Info that can *harm* data subjects, data owners, or data users if it is mishandled. Not all of it is strictly "private."
- There's a lot more of it than there used to be!
  - Increased use of computers and networks
  - Increased processing power and algorithmic knowledge
  * Decreased storage costs
- "Mishandling" can be very harmful.
  - ID theft
  - Loss of employment or insurance
  - "You already have zero privacy. Get over it."
    (Scott McNealy, 1999)

PORTIA Goals

- Produce a next generation of technology for handling sensitive information that is qualitatively better than the current generation's.
- Enable end-to-end handling of sensitive information over the course of its lifetime.
- Formulate an effective conceptual framework for policy making and philosophical inquiry into the rights and responsibilities of data subjects, data owners, and data users.

Academic-CS Participants

Stanford

Dan Boneh
Hector Garcia-Molina
John Mitchell
Rajeev Motwani

Yale

Joan Feigenbaum
Ravi Kannan
Avi Silberschatz

Univ. of NM

Stephanie Forrest
("computational immunology")

Stevens

Rebecca Wright

NYU

Helen Nissenbaum
("value-sensitive design")

Multidisciplinarity on Steroids

J. Balkin (Yale Law School)
G. Crabb (Secret Service)
C. Dwork (Microsoft)
S. Hawala (Census Bureau)
B. LaMacchia (Microsoft)
K. McCurley (IBM)
P. Miller (Yale Medical School)
J. Morris (CDT)
B. Pinkas (Hewlett Packard)
M. Rotenberg (EPIC)
A. Schäffer (NIH)
D. Schutzer (CitiGroup)

Note participation by the software industry, key user communities, advocacy organizations, and non-CS academics.

Five Major Research Themes

- Privacy-preserving data mining and privacy-preserving surveillance
- Sensitive data in P2P systems
- Policy-enforcement tools for db systems
- Identity theft and identity privacy
- Contextual integrity

ID Theft and ID Privacy

- Problem: People use the same uid/pwd at many websites.
- Example: Same uid/pwd at eBay and at a high-school alumni site
- Threat: A break-in at a low-security site reveals many uid/pwd pairs that can be used at high-security sites.

Anti-Phishing Tools

http://crypto.stanford.edu/SpoofGuard/
http://crypto.stanford.edu/PwdHash/

Students: R. Ledesma, B. Ross, and Y. Teraguchi

Faculty: D. Boneh and J. Mitchell

PwdHash is a browser plug-in that converts the user's pwd to a unique, *site-specific* pwd.

Basic Algorithm

- Locate all pwd HTML elements on page:
  <INPUT TYPE=password NAME=pass>

- When form is submitted, replace contents of pwd field with

  $$\mathbf{HMAC}_{\mathbf{pwd}}\mathbf{(domain\text{-}name).}$$

- Send pwd *hash* to site instead of pwd.

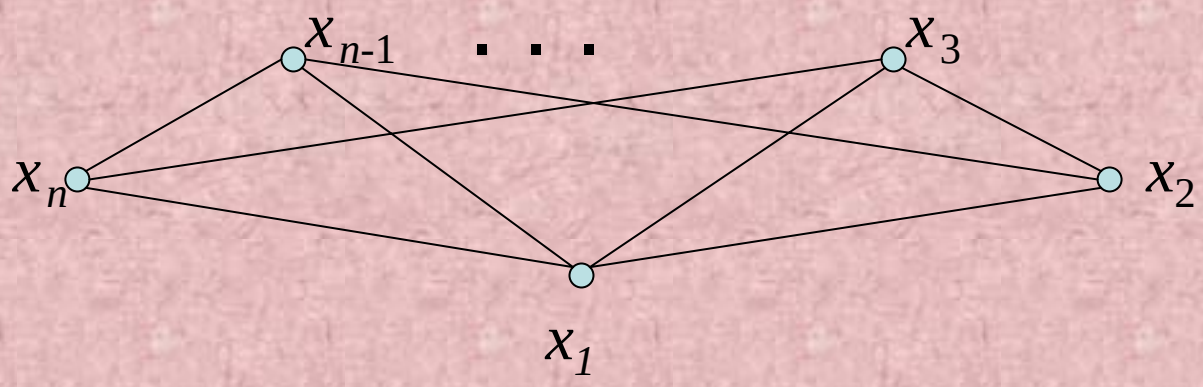
Features

- Conceptually *simple* solution!
- Implementation includes:
  - pwd-reset page
  - remote-hashing site (used in, *e.g.*, cafés)
  - list of domains for which domain of reset page is not domain of use page (*e.g.*, Passport)
- Dictionary attacks on hashes are much less effective than those on pwds and can be thwarted *globally* with a high-entropy plug-in pwd.

Privacy-preserving Data Mining

- Is this an oxymoron?
- No! Cryptographic theory is extraordinarily powerful, almost paradoxically so.
- Computing exactly one relevant fact about a distributed data set while concealing everything else is exactly what cryptographic theory enables *in principle*. But not (yet!) in practice.

Secure, Multiparty Function Evaluation



$$y = F(x_1, \ldots, x_n)$$

- Each *i* learns *y*.
- No *i* can learn anything about x_j (except what he can infer from x_i and *y*).
- Very general positive results. Not very efficient.

New Special-Purpose SMFE Protocols

- Lindell and Pinkas: Efficient 2-party protocol for ID3 data mining on $x_1 \cup x_2$
- Aggarwal, Mishra, and Pinkas: Efficient n-party protocol for order statistics of $x_1 \cup \cdots \cup x_n$
- Freedman, Nissim, and Pinkas: Efficient 2-party protocol for $x_1 \cap x_2$
- Wright and Yang: Efficient 2-party protocol for K2 Bayes-net construction on $x_1 \cup x_2$

Secure Computation of Surveys

Joan Feigenbaum (Yale), B. Pinkas (HP),

R. Ryger (Yale), and F. Saint-Jean (Yale)

http://www.cs.yale.edu/homes/jf/SMP2004.{pdf, ppt}

Surveys and other Naturally Centralized Multiparty Computations

- Consider
  - Sealed-bid auctions
  - Elections
  - Referenda
  - Surveys
- Each participant weighs the hoped-for payoffs against any revelation penalty ("loss of privacy") and is concerned that the computation be fault-free and honest.
- The implementor, in control of the central computation, must configure auxiliary payoffs and privacy assurances to encourage (honest) participation.

CRA Taulbee Survey: Computer Science Faculty Salaries

- Computer science departments in four tiers, 12 + 12 + 12 + all the rest
- Academic faculty in four ranks: full, associate, and assistant professors, and non-tenure-track teaching faculty
- Intention: Convey salary distribution statistics per tier-rank to the community at large without revealing department-specific information.

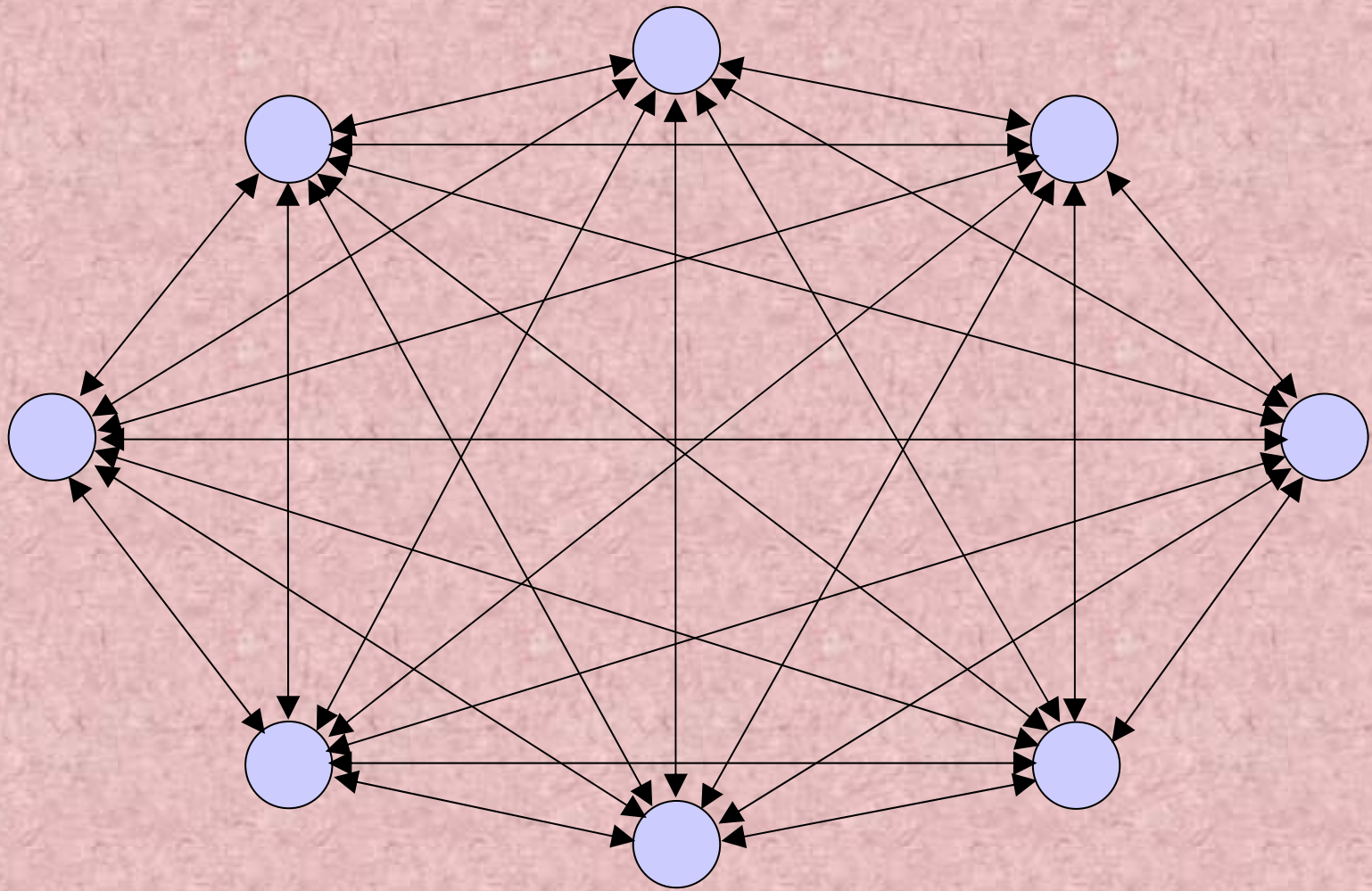
CRA Taulbee Survey: The Current Computation

- Inputs, per department and faculty rank:
  - Minimum
  - Maximum
  - Median
  - Mean
- Outputs, per tier and faculty rank:
  - Minimum, maximum, and mean of department minima
  - Minimum, maximum, and mean of department maxima
  - Median of department means (not weighted)
  - Mean (weighted mean of department means)

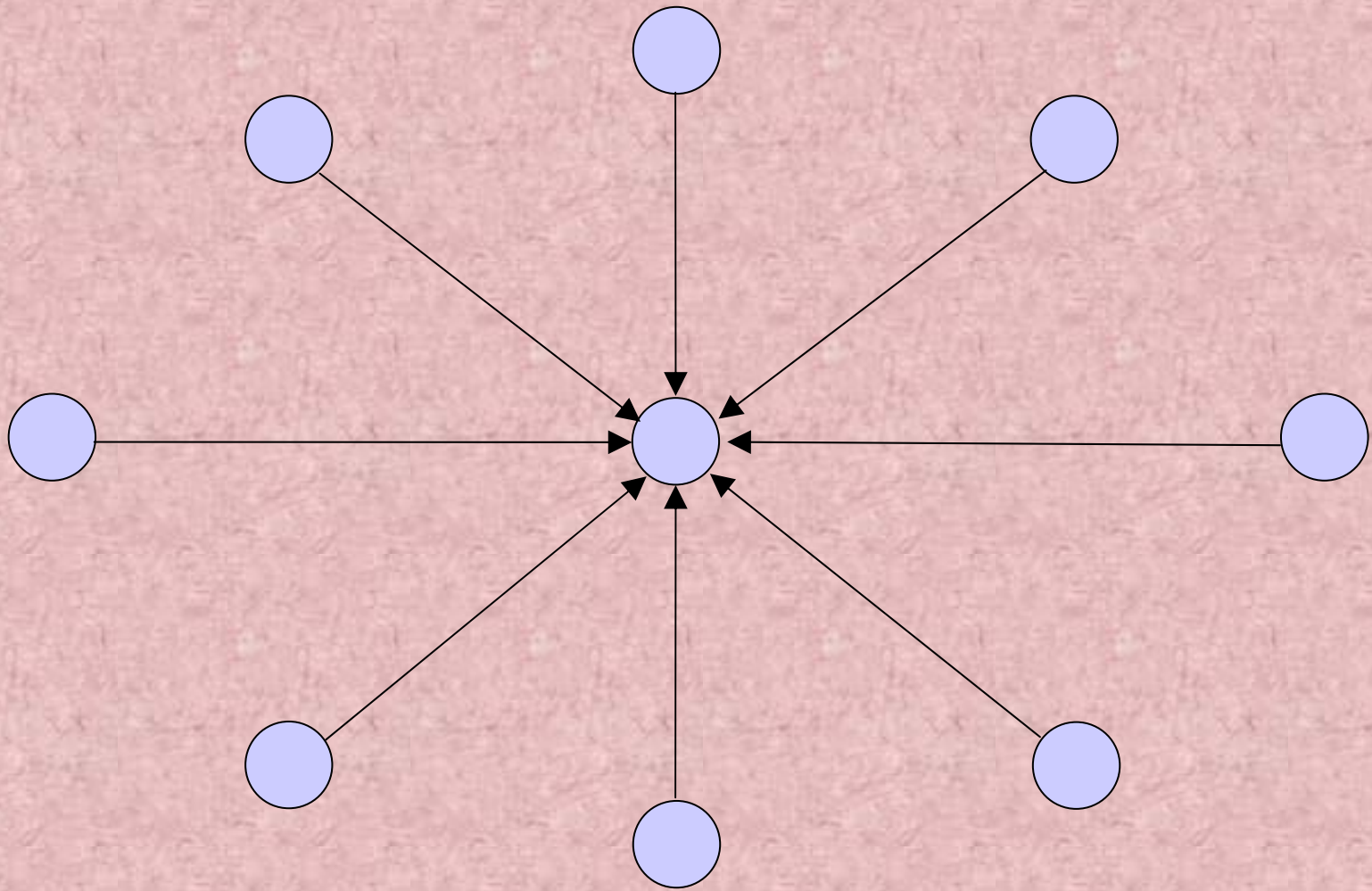
Taulbee Survey: The Problem

- CRA wishes to provide *fuller statistics* than the meager data currently collected can support.
- The current level of data collection *already compromises department-specific information*. Asking for submission of full faculty-salary information greatly raises the *threshold for trust* in CRA's intentions and its security competence.
- Detailed disclosure, even if anonymized, may be explicitly prohibited by the school.
- Hence, there is a danger of significant *non-participation* in the Taulbee Survey.

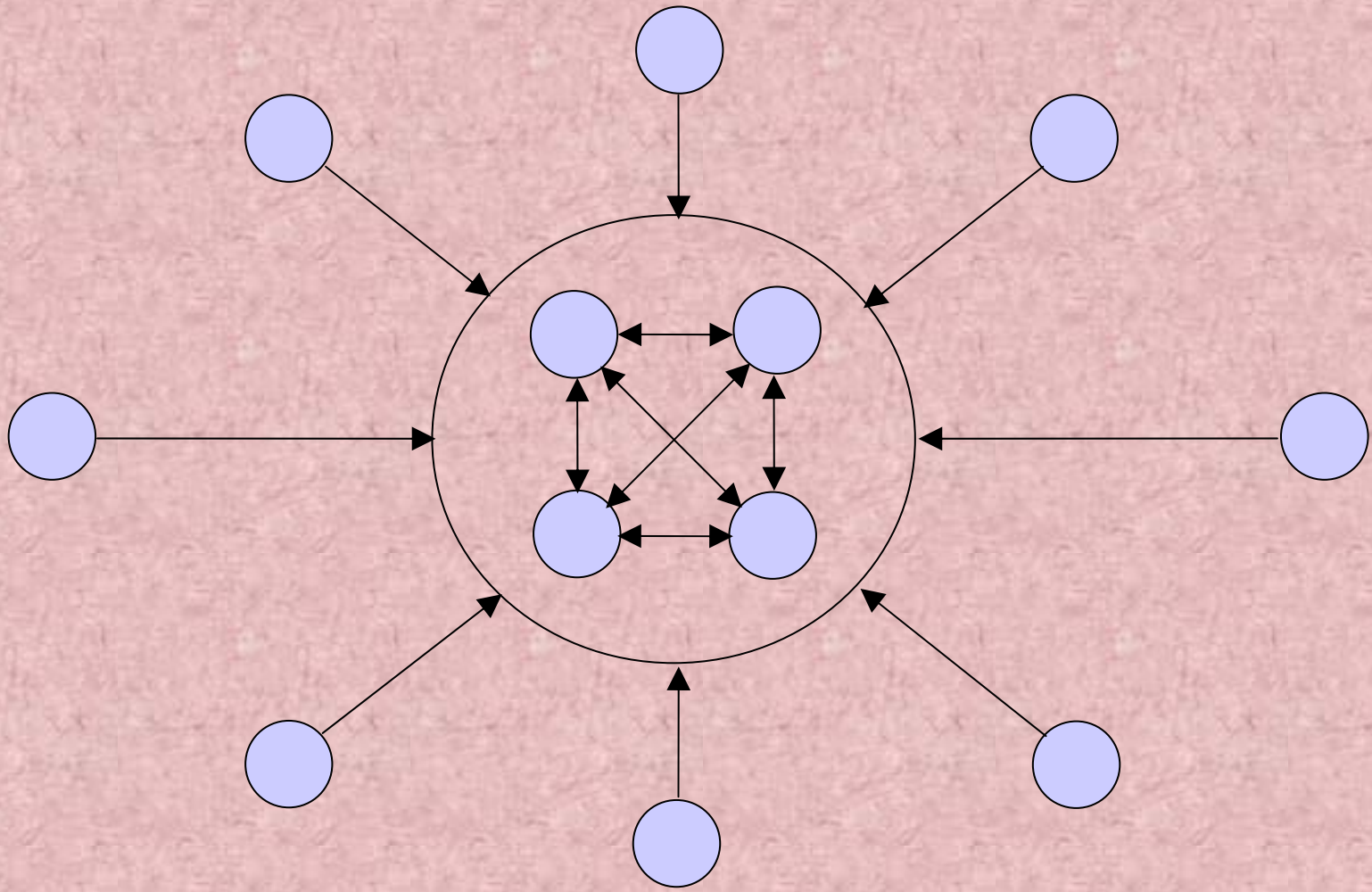
Communication Pattern: General SMFE Protocols



Communication Pattern: Surveys and Other Trusted-Party Computations



Communication Pattern: *M*-for-*N*-Party SMFE



Our Implementation: Input-Collection Phase

- Secure input collection:
  - Salary and rank *data entry* by department representative
  - Per rank, in JavaScript, *computation of XOR shares* of the individual salaries for the two (*M = 2*) computation servers
  - Per rank, *HTTPS transmission* of XOR shares to their respective computation servers
- Note that cleartext data *never leave the client machine.*

Our Implementation: Computation Phase

- Per tier and rank, *construction of a Boolean circuit* to
  - reconstruct inputs by XOR-ing their shares
  - sort the inputs in an odd-even sorting network
- Secure computation, per tier and rank:
  - *Fairplay* [Malkhi *et al.*, 2004] implementation of the *Yao 2-party SFE* protocol for the constructed circuit and the collected input shares
  - Output is a sorted list of all salaries in the tier-rank.
- Postprocessing, per tier and rank:
  - *arbitrary, statistical computation* on the sorted, cross-departmental salary list

The Heartbreak of Cryptography

- User-friendly, open-source, *free* implementation
- NO ADOPTION !@%$#
- CRA's reasons
  - ∗Need for data cleaning and multiyear comparisons
  - – Perhaps most member departments will trust us.
- Yale Provost's Office's reasons
  - ∗No *legal* basis for using this privacy-preserving protocol on data that we otherwise don't disclose
  - ∗Correctness and security claims are hard and expensive to assess, despite open-source implementation.
  - ∗All-or-none adoption by Ivy+ peer group.

PORTIA Activities also Include:

- Stream algorithms for massive graphs
  (J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, J. Zhang)
- Approximate massive-matrix computations
  (P. Drineas, R. Kannan, M. Mahoney)
- Query engines for medical databases
  (J. Corwin, P. Miller, A. Silberschatz)
- Contextual integrity
  (H. Nissenbaum)
- Legal foundations
  (J.Balkin, J. Feigenbaum, N. Kozlovski)

Stream Algorithms for Massive Graphs

- A graph with n nodes and m edges is presented as a stream of edges.
- Very little can be done when the algorithms are limited to o(n) space.
- [FKMSZ] uses n•polylog(n) space for:
  - Approximate matching
  - Approximate all-pairs shortest-path distances
- Some massive-graph problems require multiple passes in the streaming model.

Approximate Massive-Matrix Computations

- Approximate by sampling the rows and the columns of the matrices.
- Goals are fast running time and few passes over the matrices.
- [DKM] provides algorithms for:
  - Approximate matrix multiplication
  - Computing a low-rank approximation of a matrix
  - Approximating a compressed matrix decomposition

See PORTIA Website for:

- Papers, talks, and software
- Educational activities
  - Courses
  - Grad students and postdocs
- Media coverage
- Programs and slides from workshops
- Related links

[Google "PORTIA project"]

What May We Use To Prevent Unwanted Phone Calls?

+ Technology
  - Answering machines
  - Caller ID
+ Money (together with technology)
  - "Privacy-guard service" from SNET
? Government
  - "Do-Not-Call" lists seem to be controversial.

What May We Use To Prevent Unwanted Email?

+ Technology
  - Filters
  - CAPTCHAs
  - "Computational postage"
? Government
  + Yes, if the unwanted email is "trespass to chattel," which requires that it "harm" the recipient's computer system. (CyberPromotions)
  – No, if the email is merely "unwanted." (Hamidi)

Is a Network like a Country?

- Size, diversity, and universal connectivity imply risk. Get over it!
- Subnetworks ≈ neighborhoods (J Yeh, CS457)
  - Some segregation happens naturally.
  - Gov't-sanctioned segregation is wrong.
- Alternative: Network nodes ≈ homes (JF)
  - A man's computer is his castle.
  - Do I have to be rich or tech-savvy to deserve control over my own computer?

Is there a Limit to the Upside of Network Effects?

Metcalf's Law: The value to a potential user of connecting to a network grows as the square of the number of users already connected.

Feigenbaum's Law: Metcalf's Law holds only until almost all potential users, including the scum of the earth, are connected. Then the value of the network drops to zero for almost everybody.

Preliminary Conclusions

- Less and less sensitive information is truly inaccessible. The question is the *cost* of access, and that cost is decreasing.
- Foundational legal theories to support obligations and rights in cyberspace are lacking.
- Technological progress is still going strong, almost 30 years after Diffie-Hellman, but adoption is slow.

? Next step: Find a community of data owners who *need* the results of joint computations and *can't get them* without SMFE. (Medical researchers?)